1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.

- Below are the variables having any kind of trend with the target variable
    - Season
        - Summer and fall have more rentals
    - Month
        - Follows a similar trend to season
    - Yr
        - 2019 has more rental as compared to 2018
    - Weathersit
        - Clear weather results in more rental
    - Throughout the week the rentals are similar
    - Working days and holidays do not have much impact

2. Why is it important to use **drop_first=True** during dummy variable creation?

Ans.

By dropping the first columns from the dummy variable we remove one complexity from the model, without having any impact on the quality of the model

- All the categories are represented by n-1 dummy variable, n being number of categories, and the missing dummy variable category being represented by all zeroes in the rest.
- Due to the above mentioned point, the model is can be considered to be defaulted for the missing categorical dummy variable

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans.

- temp.
    - atemp also has similar corelationship

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.

- Using residual analysis.
    - The mean of error terms are zero (distribution plot for error plot)
    - Error terms are normally distributed (distribution plot for error plot)
    - Error terms are independent of each other (scatter plot between y-test and error)